Video Memorability Prediction: An investigation of the performance of Regression Models

Dao Thi Nguyen
Student ID: 20212316
DC836, MSc Computing, Data Analytics
dao.nguyen7@mail.dcu.ie

Abstract

Video memorability prediction task has recently gained a significant attention from the research community. This paper attempts to explore numerous feature fusion strategies on different regression models to predict videos' memorability scores that reflect the probability of a video being remembered to viewers. The research has been implemented based on the MediaEval 2018 dataset. A range of predicting techniques, i.e. Random Forest Regression, Linear Regression, Support Vector Regression and a Multilayer Perceptron (MLP) model are utilized along with 6 state-of-the-art features and their combinations. The results have shown promising performance in the accuracy of the Random Forest Regression model with the different combinations of 4 features (C3D, Aesthetics, HMP, and LBP) which outperforms other predicting approaches.

Keywords: datasets, neural networks, gaze detection, text tagging

1 Introduction

Thanks to the development of the Internet, it is the fact that a huge number of videos are posted on popular social network platforms such as Youtube, Facebook, Instagram and Tiktok [17]. This phenomenon could raise a problem related to processing exponentially increasing amounts of daily media content data in the internet-based platforms. In order to solve this problem, developing new methods to organise and retrieve useful information from digital data is necessary. To support the retrievement of digital content, the prediction of video memorability score has recently gained a significant attention from the research community. This study of video memorability will benefit numerous applications in various fields such as education, advertisement, information searching and retrieval, and content recommendation. In general, the target of this article is to answer the following research question:

- Can the video memorability be predicted by using their content?
- If yes, which existing approaches deliver the best performance in terms of prediction accuracy?

In this paper, six state-of-the-art pre-computed features will be used to train different regression predicting models. These features consist of video dedicated and frame-based features. Video dedicated features comprise of C3D - final classification layer of the C3D model [19] (101 features), and HMP - Histogram of Motion Patterns for each video [3] (6075 features). Frame-based features were extracted from three key-frames, i.e. the 0th, 56th and 112th frame of each video are considered, which correspond to the frame at the beginning, one-third and two-third of the video timeline, respectively. In each frame, the following feature types are extracted: LBP-local binary pattern [14], InceptionV3-output of the fc7layer of InceptionV3 deep network [18], Color Histogram-classic color histogram (three channels) and Aesthetic visual features which is a collection of features used in the prediction of visual aesthetics, composed of color, texture and object- based descriptors [9]. The research has been carried out based on the MediaEval 2018 dataset [6]. In addition, a range of machine learning algorithms (different types of Linear Regression, Random Forest Regression and Support Vector Regression algorithms and MLP models) is utilised to predict the memorability of videos. An evaluation metric, namely Spearman's rank correlation, has been used for the comparison of the performance of different models on the six state-of-the-art features.

The rest of this paper is structured as follows: Section 2 provides the literature review on video memorability predictions. Section 3 presents the understanding video memorability task. The research method of the paper and experimental results are presented and discussed in Section 4, followed by a conclusion and future work in Section 5.

2 Related Work

Several previous works have solved the memorability tasks mainly using the pre-computed features [7, 8, 10, 11, 13, 16, 20, 21]. The memorability task can be done using single or multi-features information to train a regression model. The research works on video memorability have been inspired by the success of image memorability in [11, 13] in building models to predict image memorability from visual features. In 2015, [10] trained a Support Vector Regression (SVR) model to map these features to memorability scores, estimating the memorability score of a test video clip.

In 2018, the authors in [7, 16] trained a Multi-Layer Perceptron (MLP) on top of every single feature and a combination of the three best non-image caption-based features. Meanwhile, over-fitting is potentially a primary concern in the

memorability task. To prevent over-fitting, the use of images information source via linear highly regularized models was introduced in [8], utilising the provided features consisting of Residual and Dense Network features. In this approach, the authors use Least Absolute Shrinkage and Selection Operator (LASSO), SVR, and Elastic Network (ENet) for their experiments. The result has demonstrated that Lasso Logistic Regression appear to be the best model for the HMP, LBP and ColorHistogram features, while ElasticNet has been proved to be the best model for the other features.

In 2020, the authors in [20] indicated that Random Forest (RF) Regression model has delivered the best performance in predicting the long-term memorability score using C3D fusing semantic features. Meanwhile, the SVR model has outperformed in predicting the short-term memorability score using LBP fusing semantic features.

In 2021, the author [21] proved that LBP, VGG, and C3D have demonstrated better performances in comparison with other provided features. Moreover, they also discovered that SVR models produced the best results with these features.

The approaches discussed above have demonstrated their effectiveness in video memorability prediction. In this paper, regression-based methods (Linear Regression, SVR, RF Regression and MLP) are seek to use on different pre-extracted features.

3 Understanding Video Memorability

3.1 Video Memorability Measurement Protocol

The protocol comprises of two stages and is based on the recognition tests for memorability scores. The first stage includes interlaced observation and recognition tasks. Participants watched a sequence of 180 different videos that they were not familiar with, including 40 targets video and 140 fillers. Target videos will be repeated during the participants' observation. If they recognised a repetition, a click on the space bar is required. After 1-3 days of performing the long-term memorability video stage, the participants watched a new sequence of videos including 40 targets chosen from the fillers of the first stage, and 120 new fillers. If they think that they have seen the video, they can press the space bar.

After the two stages, video memorability scores were calculated for each video based on the proportion of the right detections of target videos by participants for the memory performances in both short term and long term. That is the way ground truth has been measured through recognition test.

3.2 Memorability Dataset Description

The dataset consists of 8,000 short videos with no sound shared under a license. Participants can use and redistribute them in the context of MediaEval 2018. The dataset was divided into the development set (6000 videos) with a set of pre-extracted state-of-the-art visual features and ground truths of both short-term and long-term memorability available and test set (2000 videos) with all the features but no ground truth. Videos are varied and have different types of scenes. Each video has 7s-duration and is also attached by descriptive titles (captions).

3.3 Task Description

The task concentrates on the issue of predicting how memorable a video is to viewers. Participants are required to training computational models that have the capability of automatically predicting videos' memorability scores reflecting the probability a video will be remembered from visual content. Video captions or title descriptions attached to the videos are not allowed to be used. Spearman's rank correlation will be employed as an evaluation metric in ranking tasks.

4 Memorability Prediction Methods

4.1 Data Preprocessing

4.1.1 Image/video Features. Video features (C3D, HMP) and image features (inceptionV3, LBP, color Histogram, Aesthetics) are employed for predicting the memorability score. Generally, these pre-extracted visual features data and ground truth are loaded from the shared dataset and then merged together for further training. Since the resulting dimensional space of HMP, InceptionV3 and Color Histogram features is significant high, their dimensions are reduced based on Principal Component Analysis (PCA) technique [12]. PCA helps to reduce the number of dimensions of a dataset, by transforming it into a smaller vector space that still contains most of the variance in the original data. This transformation process facilitates to analyse the data much easier and faster by using machine learning algorithms without extraneous variables to process.

In addition, because there are missing values in the LBP feature data of the first frame of 15 videos, these missing values should be imputed to facilitate for training phase. Data of the second frame on the same videos on same feature is used to impute with.

4.1.2 Feature Fusion. Image and video features are combined via early fusion. Prior to this step, the features' dimensionality is reduced by using PCA with the support of scikit-learn library in Python. The reduced data still contains more than 95% variance of the HMP, InceptionV3, and Color Histogram features. The number of dimensions of these feature types is transformed from 6075, 3000 and 1535 into 239, 49 and 644 dimensions, respectively. This can help for better feature representation. The reduced data are then fed into the regression models.

The PCA effectiveness can also be verified before the early fusion and feature selection. While an improvement of around 2-4% in Spearman's rank correlation is observed in HMP and Color Histogram features, a degradation can

be seen in the score of other features. Based on the training results of an individual feature on different models, the four best features are chosen for feature fusion in priority, and then fed into models for further evaluation the performance of their combinations.

4.2 Model Algorithms and Training

4.2.1 Support Vector Regression Algorithm. SVR is a supervised-learning method that trains using a symmetrical loss function, which equally penalises high and low misestimation. One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space [4]. Additionally, it has excellent generalisation capability, with high prediction accuracy [4, 5].

4.2.2 Linear Regression Algorithm. The following are two main methods that are intended for regression.

- Ordinary Least Squares (OLS). It is used to fit a linear model with coefficients to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation [1].

- Regularization. The methods are extensions of the training of linear model. These seek to minimize the total of the squared error of the model on the training data (using OLS) but also to reduce the model complexity (like absolute number or size of the total of all coefficients in the model) [1] Two popular examples of regularization procedures for linear regression are:

- Lasso Regression: in which OLS is modified to also minimize the absolute total of the coefficients (L1 regularization).
- Ridge Regression: in which OLS is modified to also minimize the squared absolute total of the coefficients (L2 regularization).
- Elastic Network: is a linear regression model trained with both 1 and 2-norm regularization of the coefficients. This combination allows to learn a sparse model where some weights are nonzero like Lasso, while maintaining the Ridge's regularization properties. The convex combination of 1 and 2 is controlled by using the l1 ratio parameter. Elastic Network is useful when there are multiple features that are correlated with one another. Lasso has the ability to randomly select one of these, while elastic-net is likely to select both. One practical benefit of trading-off between Lasso and Ridge is that it allows Elastic Network to inherit some of Ridge's rotational stability.

4.2.3 Random Forest Algorithm. Random Forest [15] is an algorithm that integrates multiple trees through the idea of ensemble learning. It is also a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

4.2.4 Multi-Layer Perceptrons (MLP). MLP refers to networks composed of multiple layers of perceptrons (with threshold activation). It is commonly used in simple regression problems. MLP uses parameter alpha for regularization (L2 regularization) term which helps in avoiding overfitting by penalizing weights with large magnitudes.

4.3 Evaluation Metric

The choice of the Spearman's rank correlation as an official measure indeed corresponds to a desire of normalizing the output of the different systems and making the comparison easier [6].

Spearman's coefficient is used for the predicted video memorability score and ground truth. When > 0, it means that the video memorability score is positively correlated with ground truth, and the closer to 1, the better the result.

5 Experimental Results

All below experimental results are collected from the models that are trained with a training set (80% development dataset - 4800 videos) and validated with a test set (20% development dataset -1200 videos) thanks to the support of Scikit-Learn and Keras library in Python. The sourcecode of the experiments in this paper can be found in [2].

5.1 Single Feature Prediction

C3D, HMP, LBP, InceptionV3, Color Histogram, and Aesthetic features are employed as input, and Linear Regression, Random Forest, SVR and MLP are used for training the predicting models. Overall, the RF Regression model outperforms the other models in predicting either short-term or long-term memorability in most of the feature types.

It can be seen from Table 1 that, for short-term video memorability prediction, the Spearman correlation coefficient obtained by using the random forest algorithm to train the input C3D feature is better than the results for other models and features.

For long-term video memorability, the results obtained by using the random forest algorithm with input Aesthetic features are better than those obtained by other features, as can be seen from Table 2.

5.2 Multi- feature Prediction

According to the results of single-feature training, four features along with models that have the best training results are selected for multi-feature fusion and analyzed whether the multi-feature fusion has a better effect on predicting video memorability. The results are shown in Table 3 and 4.

Through the comparison of the two-level training results of a single feature, multi-feature, it can be seen that, in general, the results obtained after training the data by fusion of

Table 1. Training results of single feature- Spearman's ranking correlation for Short-term memorability

Model/Feature	*C3D*	*HMP*	*LBP*	*ColorHis*	*InceptionV3*	*Aesthetics*
Linear Regression	0.277	0.036	0.177	0.045	0.094	0.283
Ridge	0.2837	0.208	0.190	0.044	0.157	0.272
Lasso	0.280	0.297	0.254	0.044	0.143	0.280
Elastic Network	0.279	0.182	0.178	0.044	0.161	0.230
SVR	0.283	0.296	0.220	0.252	0.156	0.182
RF Regression	**0.327**	**0.298**	**0.296**	0.290	0.120	**0.318**
MLP	0.277	0.288	0.216	0.162	0.170	0.158

Table 2. Training results of single feature - Spearman's ranking correlation for Long-term memorability

Model/Feature	*C3D*	*HMP*	*LBP*	*ColorHis*	*InceptionV3*	*Aesthetics*
Linear Regression	0.103	0.036	0.048	0.024	0.052	0.124
Ridge	0.123	0.090	0.066	0.024	0.093	0.118
Lasso	**0.132**	0.117	0.058	0.029	0.087	0.123
Elastic Network	0.127	0.078	0.059	0.029	0.091	0.090
SVR	0.063	0.106	0.079	0.082	0.055	0.087
RF Regression	0.120	0.121	**0.103**	0.102	0.013	**0.133**
MLP	0.102	**0.130**	0.071	0.062	0.088	0.094

Table 3. Training results of multi-feature, Spearman's ranking correlation for Short-term memory

Feature Fusion/Model	*RF Regression*	*Lasso*	*MLP*
C3D +Aesthetic	0.339	0.314	0.165
C3D +Aesthetic +HMP pca	0.342	0.314	0.144
C3D +Aesthetic +LBP	**0.367**	0.317	0.144

Table 4. Training results of multi-feature, Spearman's ranking correlation for Long-term memory

Feature Fusion/Model	*RF Regression*	*Lasso*	*MLP*
C3D +Aesthetic	**0.146**	0.135	0.070
C3D +Aesthetic +HMP pca	0.140	0.137	0.040
C3D +Aesthetic +LBP	0.129	0.136	0.059

Table 5. Training results on RF Regression using the features concatenated from the *Based features* ***(BF)***, i.e. including C3D, Aesthetics, LBP and the below features:

Combinations	*BF+ HMP pca*	*BF+ HMP pca+ ColorHis*	*BF + HMP pca+ ColorHis + InceptionV3*
Short-term	0.363	0.359	0.351
Long-term	0.140	0.126	0.126

features on the RF regression are better than the results for other models.

For long-term video memorability prediction, using the feature obtained by concatenating C3D and Aesthetic features to train the random forest regression model brings the best result, with 0.146 Spearman's score 4. Meanwhile, utilising the feature obtained by concatenating C3D, Aesthetics and LBP features to train the RF regression achieves the best performance with 0.367 Spearman's score for short-term video memorability prediction 3.

In comparison with single features, the concatenate features can promote the improvement of video memorability prediction results to a certain extent. After a multi-feature fusion of semantic features, the prediction results decreased 5. We notice that the feature dimension became larger after multi-feature fusion semantics. After the PCA method was used to reduce the dimensionality, and the extraction of principal components was insufficient, resulting in data loss.

Based on the above results, we suggest that the C3D feature fusing Aesthetic features yield the best performance with the RF model for the long-term and the combinations of C3D, Aesthetic and LBP features yield the best performance with the RF model for the short-term memorability scores.

6 Conclusion and Future Work

Predicting video memorability has been considered as an useful task in visual data organization, digital content search and retrieval. In this paper, we conducted a number of experiments related to the prediction of memorability scores of videos. In particular, different regression algorithms are

utilised to train predicting models, using six state-of-the-art pre-extracted features collected from the MediaEval 2018 dataset. The predicting results are evaluated by using Spearman's ranking correlation. The experimental results deliver the following findings:

- RF Regression model appears to be the best model in predicting short-term memorability using the combination of C3D, Aesthetics and LBP features and long-term memorability using C3D and Aesthetic feature concatenation. The highest Spearman's scores on training models are 0.367 for short-term memorability and 0.146 for long-term memorability prediction.
- Training results have seen a gradual decrease after concatenating more PCA-preprocessing features to the based features (C3D, Aesthetics and LBP).
- C3D appears to be the most valuable single feature in predicting video memorability, followed by Aesthetic feature.

In the future work, exploring more image or video features such as emotion along with other models to predict video memorability would be considered to improve the Spearman's ranking correlation.

References

[1] 2021. Linear Model. https://scikit-learn.org/stable/modules/linear_model.html

[2] 2021. Source Code. https://github.com/dao-nguyen0912/CA684_predicting_video_memorability

[3] Jurandy Almeida, Neucimar J Leite, and Ricardo da S Torres. 2011. Comparison of video sequences with histograms of motion patterns. In *2011 18th IEEE International Conference on Image Processing*. IEEE, 3673–3676.

[4] Mariette Awad and Rahul Khanna. 2015. Support vector regression. In *Efficient learning machines*. Springer, 67–80.

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 1–27.

[6] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. 2018. Mediaeval 2018: Predicting media memorability task. *arXiv preprint arXiv:1807.01052* (2018).

[7] Romain Cohendet, Claire-Hélène Demarty, and Ngoc QK Duong. 2018. Transfer Learning for Video Memorability Prediction.. In *MediaEval*.

[8] Rohit Gupta and Kush Motwani. 2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features.. In *MediaEval*.

[9] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, et al. 2015. Can we measure beauty? Computational evaluation of coral reef aesthetics. *PeerJ* 3 (2015), e1390.

[10] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. 2014. Learning computational models of video memorability from fMRI brain imaging. *IEEE transactions on cybernetics* 45, 8 (2014), 1692–1703.

[11] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *CVPR 2011*. IEEE, 145–152.

[12] Ian T Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2065 (2016), 20150202.

[13] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*. 2390–2398.

[14] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence* 24, 7 (2002), 971–987.

[15] Christian Robert. 2014. Machine learning, a probabilistic perspective.

[16] Hammad Squalli-Houssaini, Ngoc QK Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. 2018. Deep learning for predicting image memorability. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2371–2375.

[17] Julian Struck, Fabian Siegel, Mario Kramer, Igor Tsaur, Axel Heidenreich, Axel Haferkamp, Axel Merseburger, Hendrik Borgmann, and Johannes Salem. 2018. PD23-12 UTILIZATION OF FACEBOOK, TWITTER, YOUTUBE AND INSTAGRAM IN THE PROSTATE CANCER COMMUNITY. *The Journal of Urology* 199, 4S (2018), e484–e485.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[20] Fumei Yue, Jing Li, and Jiande Sun. [n.d.]. Insights of Feature Fusion for Video Memorability Prediction. In *Digital TV and Wireless Multimedia Communication: 17th International Forum, IFTC 2020, Shanghai, China, December 2, 2020, Revised Selected Papers*. Springer Nature, 239.

[21] Tony Zhao, Irving Fang, Jeffrey Kim, and Gerald Friedland. 2021. Multimodal Ensemble Models for Predicting Video Memorability. *arXiv preprint arXiv:2102.01173* (2021).

Declaration on Plagiarism

I declare that this material, which we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. we understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should we engage in plagiarism, collusion or copying. We have read and understood the Assignment Regulations. We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines.

Name: Dao Thi Nguyen

Date: 27/04/2021